ISPs and web privacy: insights from research

Arvind Narayanan and Dillon Reisman
Princeton Center for Information Technology Policy

Ex-parte comments to FCC, June 2016

Who we are

Arvind Narayanan, Assistant Professor of Computer Science, Princeton

Prior work: De-anonymization of Netflix Prize dataset
“Do Not Track” standard
Textbook on Bitcoin and Cryptocurrency Technologies

Dillon Reisman, Research Engineer, Princeton

Prior work: Google Privacy Team

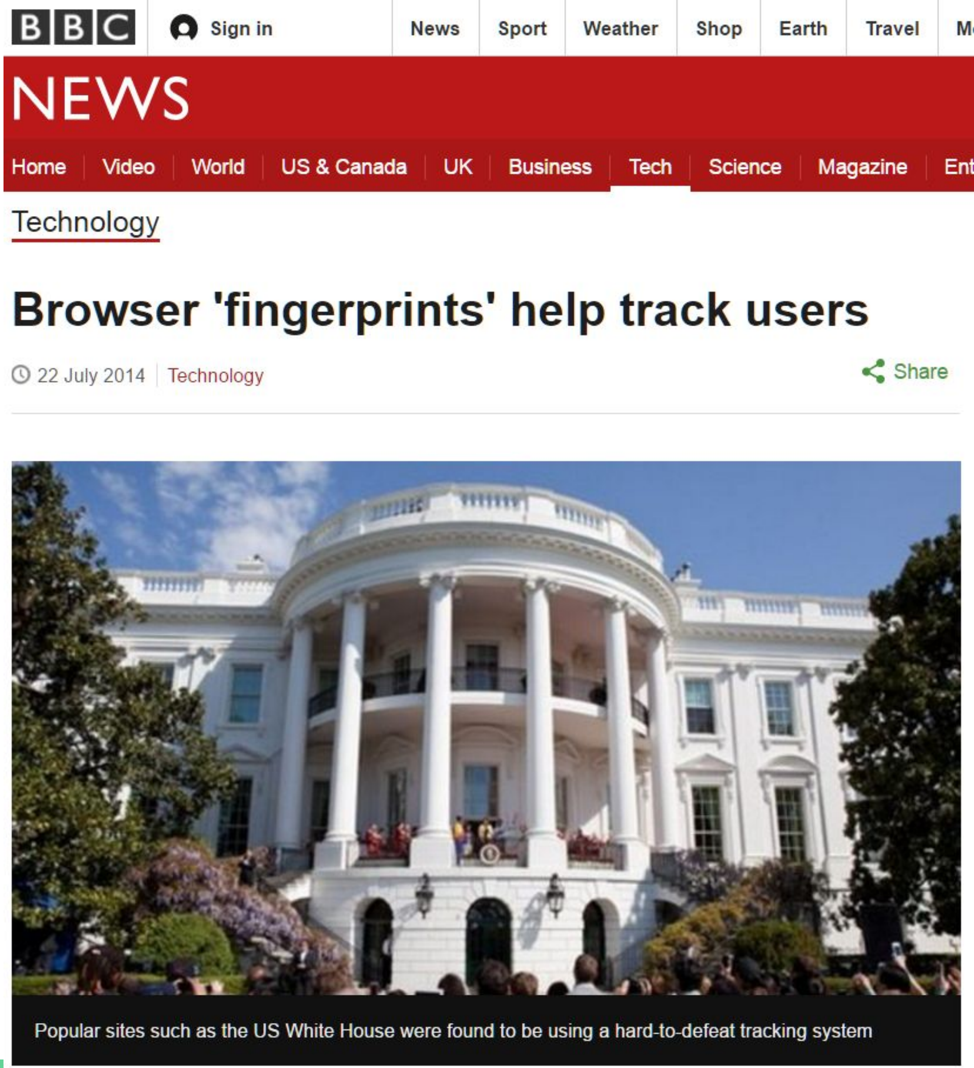
CENTER FOR INFORMATION TECHNOLOGY POLICY
AT PRINCETON UNIVERSITY

What we do

Web Transparency and Accountability Project

Monthly “privacy census” of the top **1 million** websites.

Several high-profile discoveries of hidden online tracking mechanisms.



Technology

Browser 'fingerprints' help track users

22 July 2014 | Technology

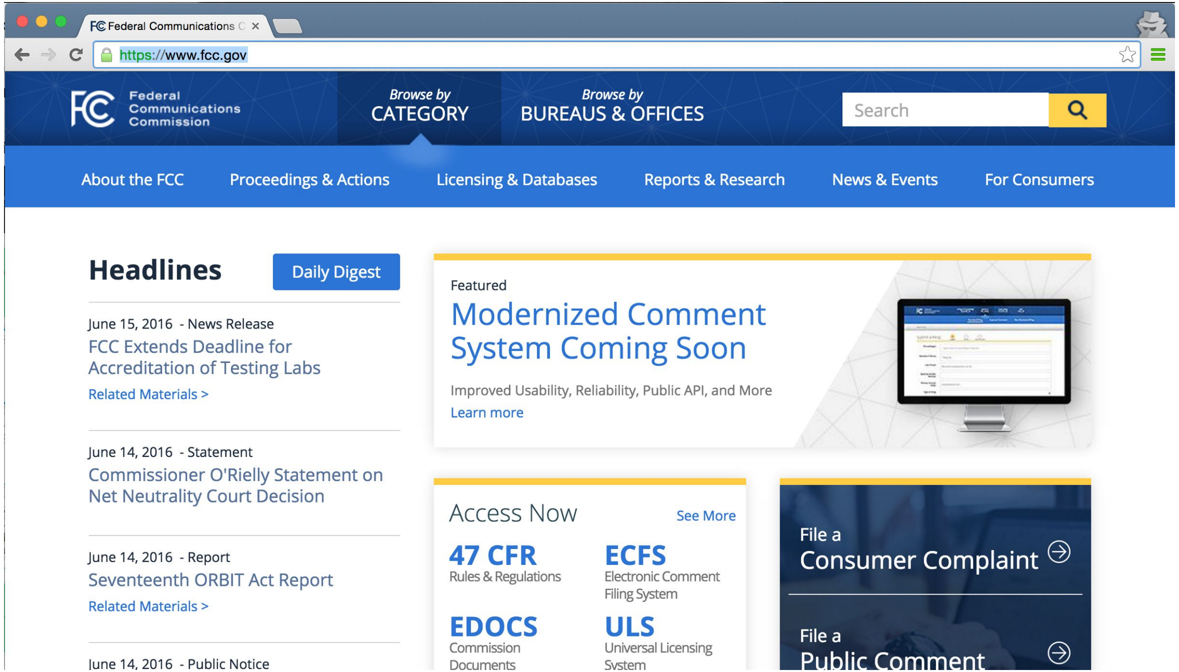
Popular sites such as the US White House were found to be using a hard-to-defeat tracking system

The Web is a mess

What happens when you visit www.fcc.gov

A visit to fcc.gov...

- 51 separate requests are made for resources on the page.
- 7 different third-party domains (not controlled by the FCC) provide resources, including:
  - Google
  - Doubleclick
  - Twitter



Headlines

Daily Digest

June 15, 2016 - News Release
FCC Extends Deadline for Accreditation of Testing Labs
Related Materials >

June 14, 2016 - Statement
Commissioner O'Rielly Statement on Net Neutrality Court Decision

June 14, 2016 - Report
Seventeenth ORBIT Act Report
Related Materials >

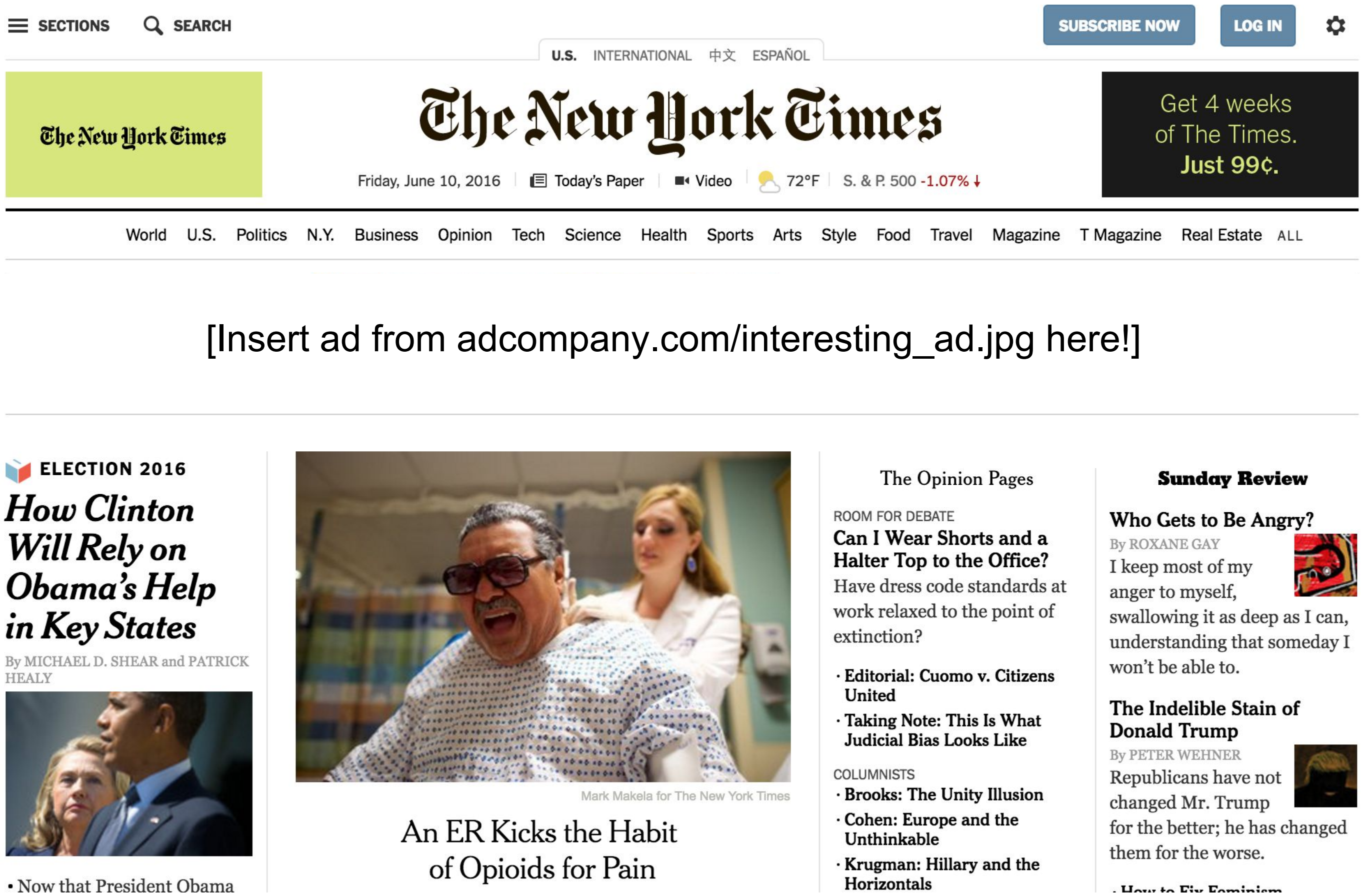
June 14, 2016 - Public Notice

Featured
Modernized Comment System Coming Soon
Improved Usability, Reliability, Public API, and More
Learn more

Access Now
See More
47 CFR
Rules & Regulations
ECFS
Electronic Comment Filing System
EDOCS
Commission Documents
ULS
Universal Licensing System

File a Consumer Complaint

File a Public Comment



SECTIONS
SEARCH
SUBSCRIBE NOW
LOG IN
U.S. INTERNATIONAL 中文 ESPAÑOL

The New York Times

Friday, June 10, 2016 | Today's Paper | Video | 72°F | S. & P. 500 -1.07% ↓

World U.S. Politics N.Y. Business Opinion Tech Science Health Sports Arts Style Food Travel Magazine T Magazine Real Estate ALL

[Insert ad from adcompany.com/interesting_ad.jpg here!]

ELECTION 2016

How Clinton Will Rely on Obama's Help in Key States

By MICHAEL D. SHEAR and PATRICK HEALY

- Now that President Obama

Mark Makela for The New York Times

An ER Kicks the Habit of Opioids for Pain

The Opinion Pages

ROOM FOR DEBATE

Can I Wear Shorts and a Halter Top to the Office?

Have dress code standards at work relaxed to the point of extinction?

- **Editorial: Cuomo v. Citizens United**
- **Taking Note: This Is What Judicial Bias Looks Like**

COLUMNISTS

- **Brooks: The Unity Illusion**
- **Cohen: Europe and the Unthinkable**
- **Krugman: Hillary and the Horizontals**

Sunday Review

Who Gets to Be Angry?

By ROXANE GAY

I keep most of my anger to myself, swallowing it as deep as I can, understanding that someday I won't be able to.

The Indelible Stain of Donald Trump

By PETER WEHNER

Republicans have not changed Mr. Trump for the better; he has changed them for the worse.

- How to Fix Feminism

nytimes.com makes a total of **195 requests,** many of them to third parties

To talk about encryption and privacy, we must understand what’s on the web.

What we’d like to discuss today

1. How much of the web is encrypted?
2. What’s visible to ISPs on encrypted and unencrypted traffic?
3. What can an ISP *infer* that is not directly revealed?
4. How effective is de-identification at protecting privacy?

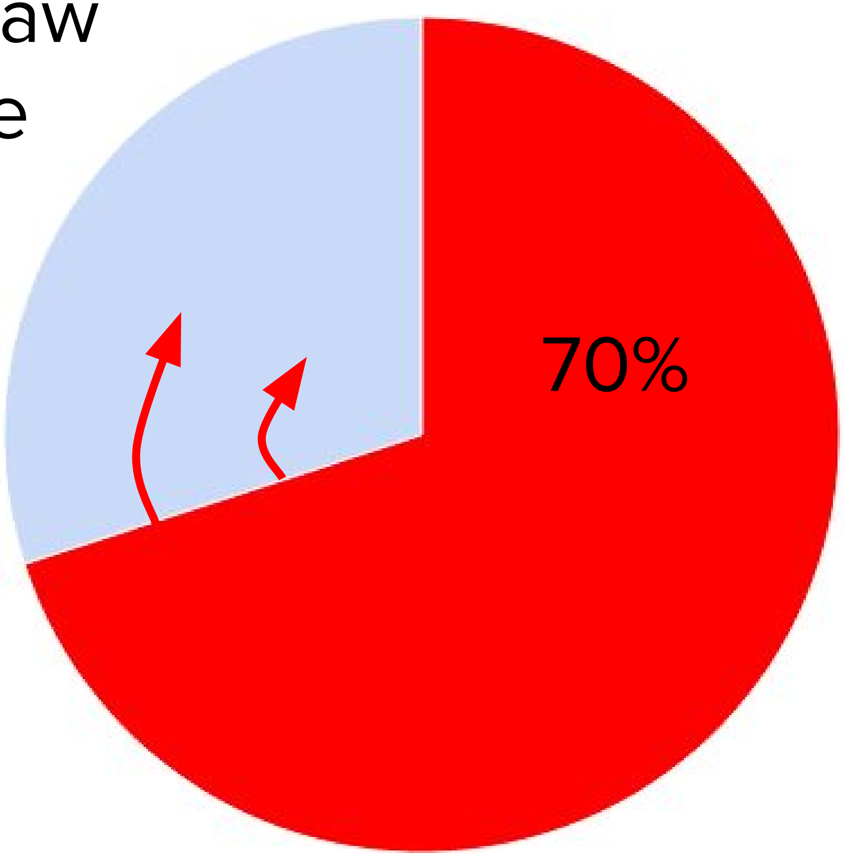
The prevalence of web encryption remains low

From Swire, et al.*:

“...Based on analysis of one source of Internet backbone data, the HTTPS portion of total traffic has risen from 13 percent to 49 percent just since April 2014. **An estimated 70 percent of traffic will be encrypted by the end of 2016.** Encryption such as HTTPS blocks ISPs from having the ability to see users’ content and detailed URLs. There clearly can be no “comprehensive” ISP visibility into user activity when ISPs are blocked from a growing majority of user activity.”

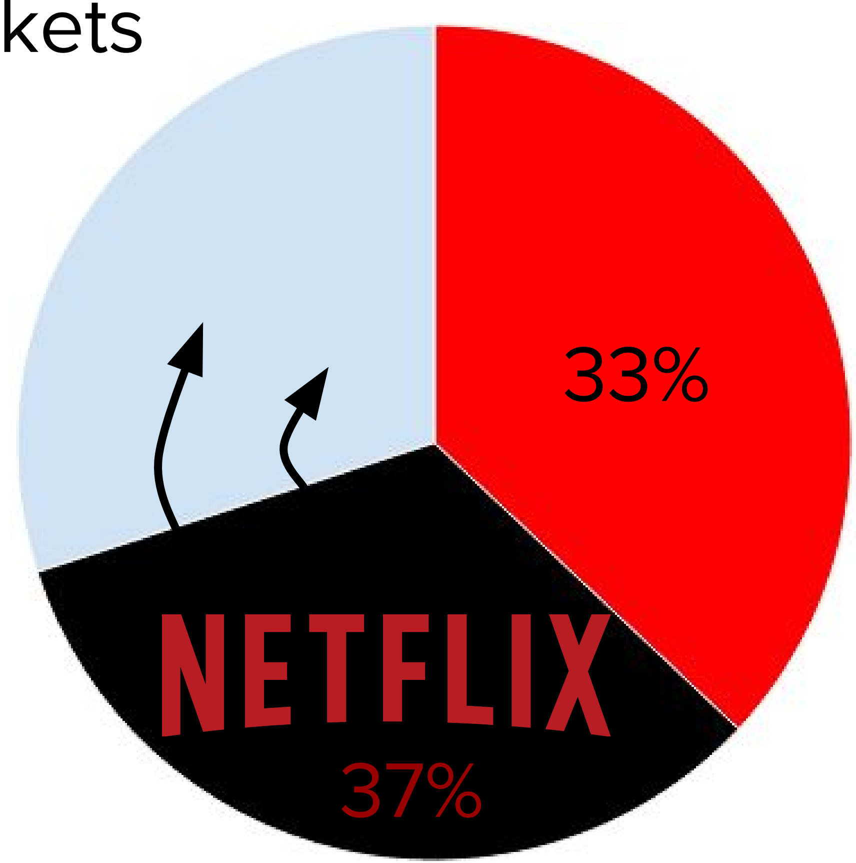
* Swire, et al., Online Privacy and ISPs: ISP Access to Consumer Data is Limited and Often Less than Access by Others (2016) http://www.iisp.gatech.edu/sites/default/files/images/online_privacy_and_isps.pdf

A majority of raw packets on the web are encrypted…



Share of web traffic served over HTTPS (encrypted)

…but raw packets make for a misleading metric.



Share of web traffic served over HTTPS (encrypted)

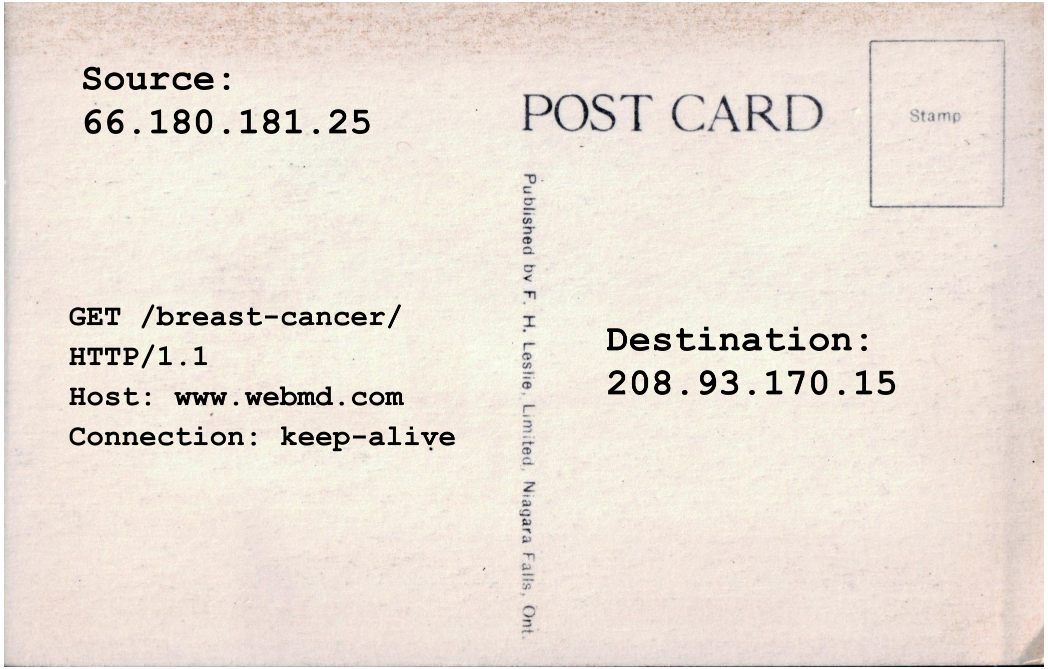
Raw traffic statistics are not meaningful in this debate.

- Streaming video services on the web account for 70% of web traffic total, but this could bear no relation to how much streaming users actually do. [1]

Bank of America
ESPN POLITICO reddit
ZocDoc WebMD℠ amazon.com®
YouTube vimeo The New York Times
NETFLIX hulu

[1]Emil Protalinski,"Streaming services now account for over 70% of peak traffic in North America, Netflix dominates with 37%" (2015) http://venturebeat.com/2015/12/07/streaming-services-now-account-for-over-70-of-peak-traffic-in-north-america-netflix-dominates-with-37/

An HTTP request for a WebMD page results in a relatively small unencrypted response.



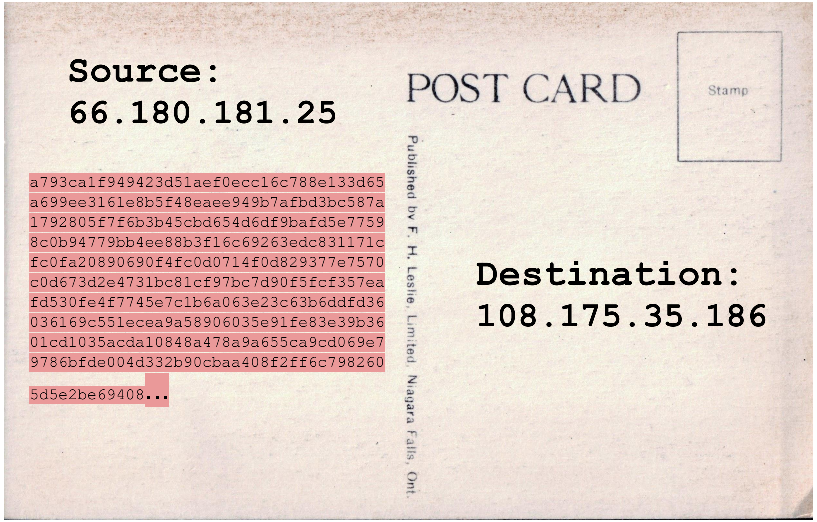
Source:
66.180.181.25

```
GET /breast-cancer/
HTTP/1.1
Host: www.webmd.com
Connection: keep-alive
```

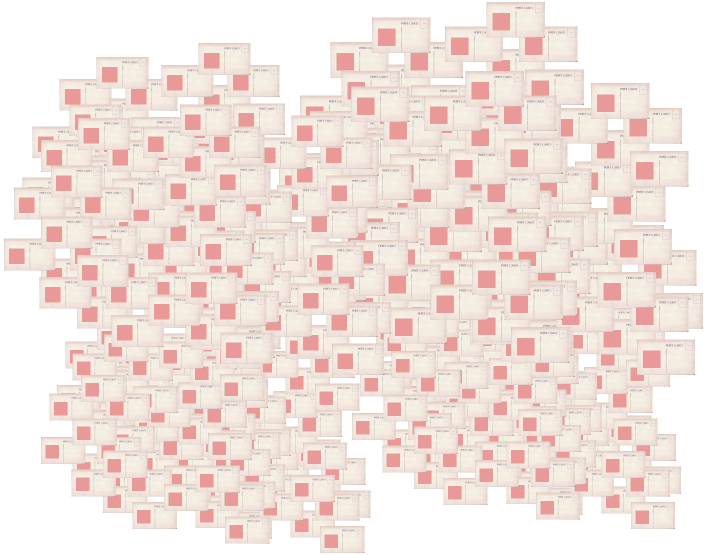
Destination:
208.93.170.15

- The application headers, and the content of the response, will be visible to the ISP.
- The main pageload for webmd.com/breast-cancer/ is 38 kB.
- On the wire, this requires **28 packets** from the server.

An HTTPS request for a Netflix HD movie results in a large encrypted response


Source:
66.180.181.25
POST CARD
Stamp
a793ca1f949423d51aef0ecc16c788e133d65
a699ee3161e8b5f48eaee949b7afbd3bc587a
1792805f7f6b3b45cbd654d6df9bafd5e7759
8c0b94779bb4ee88b3f16c69263edc831171c
fc0fa20890690f4fc0d0714f0d829377e7570
c0d673d2e4731bc81cf97bc7d90f5fcf357ea
fd530fe4f7745e7c1b6a063e23c63b6ddfd36
036169c551ecea9a58906035e91fe83e39b36
01cd1035acda10848a478a9a655ca9cd069e7
9786bfde004d332b90cbaa408f2ff6c798260
5d5e2be69408...
Published bv F. H. Leslie, Limited, Niagara Falls, Ont.
Destination:
108.175.35.186

A browser's request for a Netflix movie

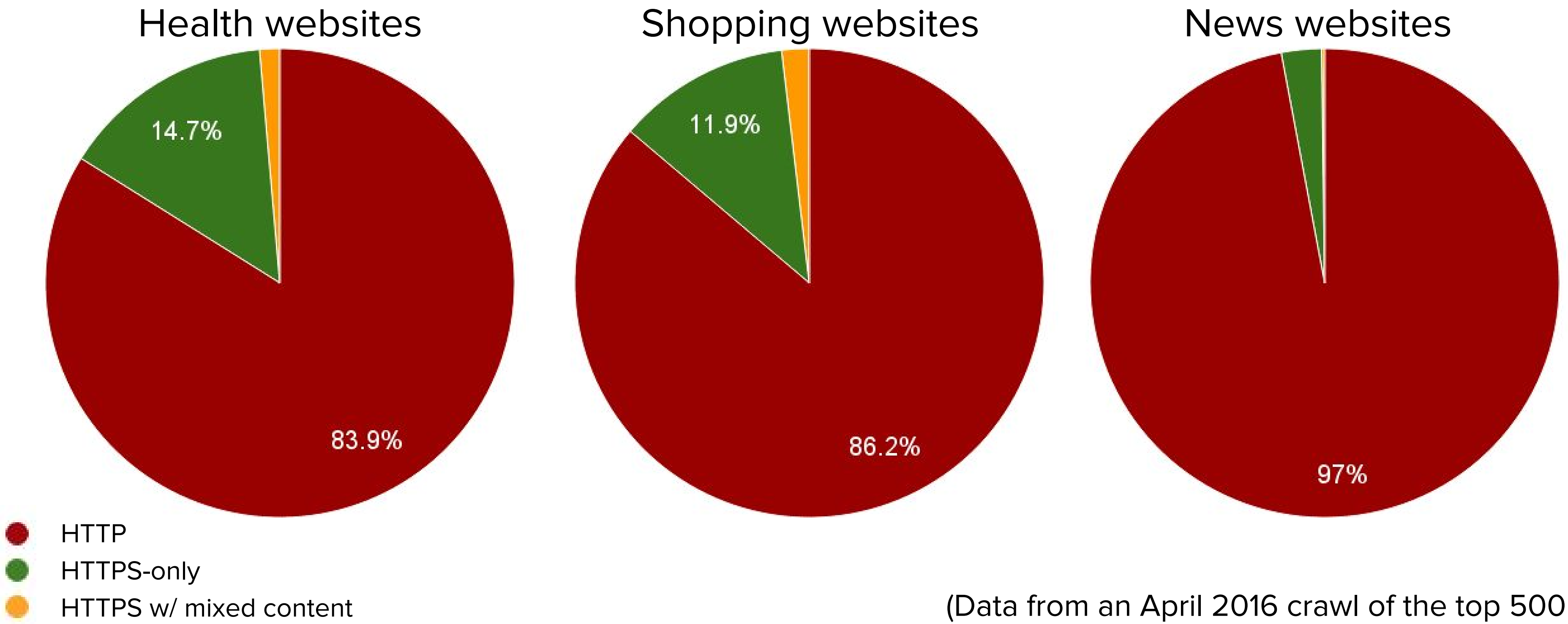


Netflix's response

A 3 GB HD movie from Netflix, assuming a packet size of 1426 bytes, will take roughly **2.1 million** packets.

What is the actual state of HTTPS on the web?

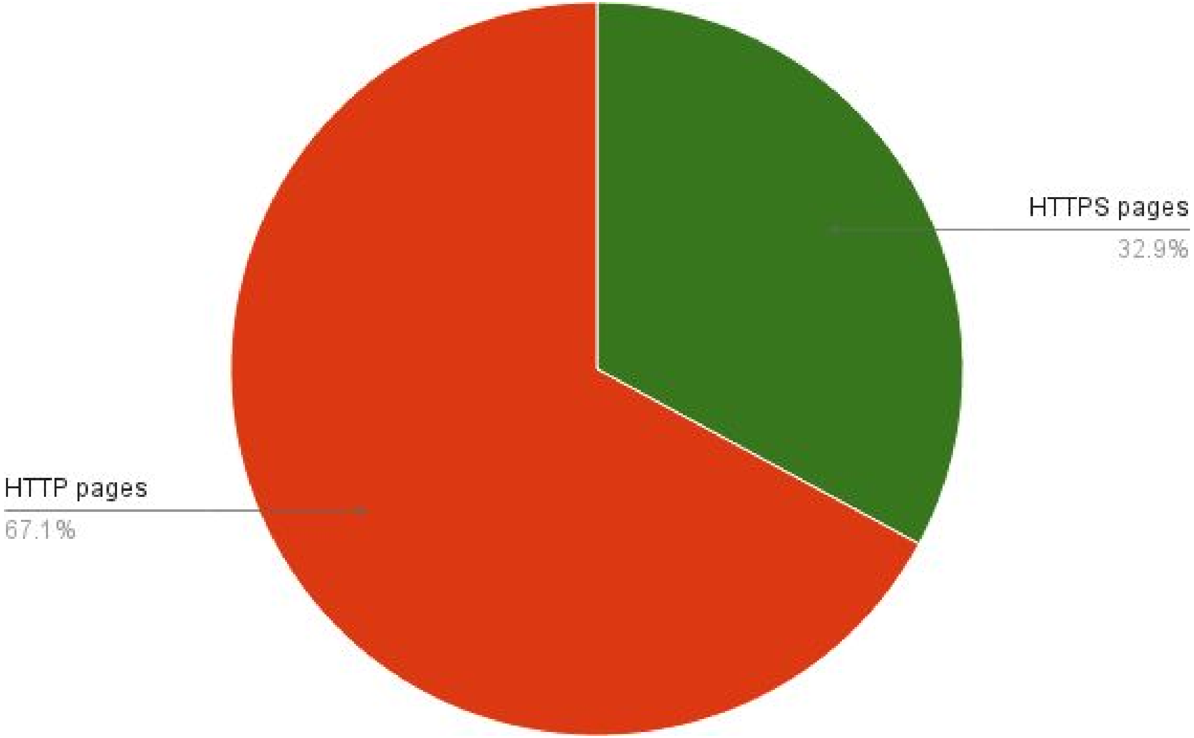
Across all categories of websites, HTTPS has a long way to go



(Data from an April 2016 crawl of the top 500 sites in each website category)

Twitter links provide a useful model for how often the average user encounters encryption

- In a sample of 1538 external links from found in public tweets in June 2016...
  - 503 links are to HTTPS pages.
  - **1025 links are to HTTP pages.**



ISP access to user information is comprehensive

What application headers look like in the web context

```
GET /2016/06/11/us/politics/hillary-clinton-obama.html?
hp&action=click&pgtype=Homepage&clickSource=story-heading&module=first-column-
region&region=top-news&WT.nav=top-news HTTP/1.1
Host: www.nytimes.com
Connection: keep-alive
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8
Upgrade-Insecure-Requests: 1a
User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_10_5) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/50.0.2661.102 Safari/537.36
Referer: http://www.nytimes.com/
Accept-Encoding: gzip, deflate, sdch
Accept-Language: en-US,en;q=0.8
Cookie: optimizelyEndUserId=oeu1461178073724r0.5984431510307393; _cb_ls=1;
```

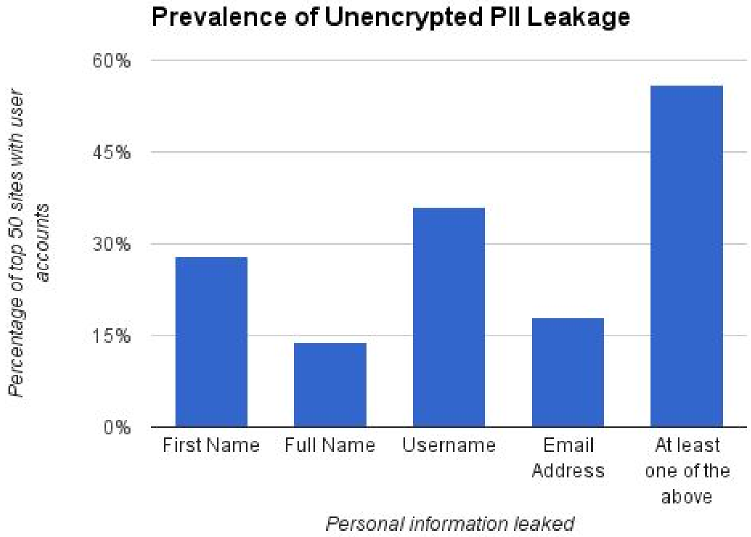
An unencrypted URL can be revealing for sensitive categories of sites

- www.webmd.com is the 2nd most popular health website online. It does not offer HTTPS, so URLs and other application headers are completely visible to ISPs
- These application headers can be both very sensitive for the end user, and potentially very valuable to an ISP
    - Example: http://www.webmd.com/lung/**mesothelioma-tests-diagnosis-and-treatments**
    - "...13 of the Top 20 most expensive keywords in 2014 were, in fact, related to mesothelioma..."[1]

[1] Jim Leichenko, "The Most Expensive Keywords in Paid Search, By Cost Per Click & Spend," 2015 https://www.adgooroo.com/resources/blog/the-most-expensive-keywords-in-paid-search-by-cost-per-click-and-ad-spend/

Unencrypted web traffic can contain customer names and other personal identifiers

- From our 2014 study[1], over **50%** of the top 50 US sites that support account creation leaked some form of personal identifier in unencrypted form.
- Leakages like these can associate traffic from one user's multiple devices.



[1] Englehardt, et al., Cookies That Give You Away: The Surveillance Implications of Web Tracking (2015) http://www.www2015.it/documents/proceedings/proceedings/p289.pdf

PII leakage happens not only on the web, but on mobile apps and IoT devices as well.

- A 2015 study[1] of 7 fitness trackers (e.g. Fitbit) and their corresponding mobile apps and web portals found at least three companies transmitted unencrypted PII.

[1] Greenwald, M., A Comprehensive Privacy Analysis of Fitness Tracker Companies (2015)

Swire, et al. on ISP visibility into full URLs[1]

> “With encrypted content, ISPs cannot see detailed URLs and content even if they try.”

* Swire, et al., Online Privacy and ISPs: ISP Access to Consumer Data is Limited and Often Less than Access by Others (2016) http://www.iisp.gatech.edu/sites/default/files/images/online_privacy_and_isps.pdf

Even when a website is encrypted via HTTPS, traffic analysis can reveal the complete URLs

- Past research in the literature has revealed that studying even encrypted web traffic can reveal sensitive information about the user.
- One method employed by UC Berkeley researchers can identify individual pages within an encrypted website with **90% accuracy**.[1]
- A study from the University of Cambridge used the amount of data transmitted over encrypted connections to infer a majority of subpages on encrypted news sites.[2]

[1] Miller, et al., I Know Why You Went to the Clinic: Risks and Realization of HTTPS Traffic Analysis (2014) https://www.petsymposium.org/2014/papers/Miller.pdf
[2] Danezis, Traffic Analysis of the HTTP Protocol over TLS http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.92.3893&rep=rep1&type=pdf

“What is the extent to which adoption of encryption technology would mitigate privacy concerns regarding broadband provider use of [deep packet inspection]. What types of information that may be learned by BIAS providers’ use of DPI are encrypted, and what types are not encrypted?”

Encryption isn’t yet widespread enough to seriously mitigate privacy concerns.

ISPs do not need application headers to provide broadband service

Application headers and packet content exist for the benefit of the web server and client.

Networks are designed so ISPs can be completely agnostic to what they contain.

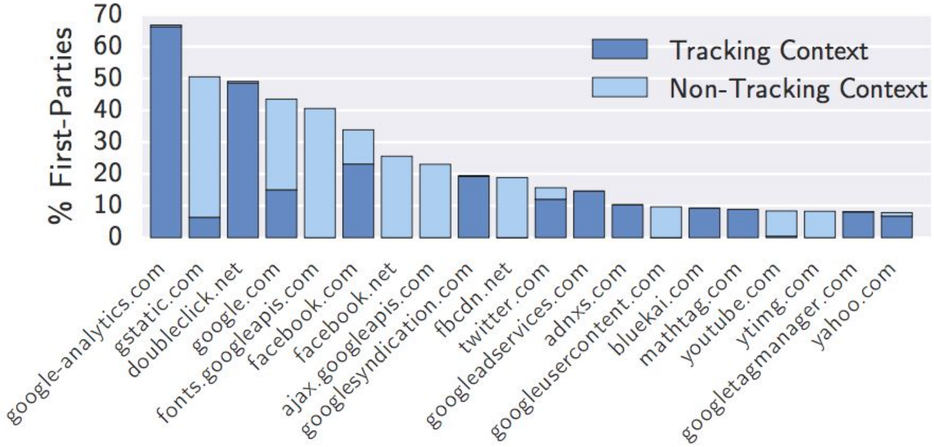
ISPs have unique — and in ways greater — visibility into a customer’s web browsing compared to non-ISP web companies

What are third parties on the web?

Third parties on the web are any resources (images, tracking pixels, advertisements, code, etc.) loaded on a webpage that come from domains that are not the main domain you visited.

Web third-parties’ visibility is limited by which first party websites include them on their pages

- Google, Facebook, and Twitter are the only third-parties present on more than 10% of the top 1 million sites.[1]
  - The most popular - Google Analytics - is not in the advertising space.
- ISPs see 100% of unencrypted web page URLs.



[1] Englehardt, S. and Narayanan, N., Online tracking: A 1-million-site measurement and analysis (2016), http://randomwalker.info/publications/OpenWPM_1_million_site_tracking_measurement.pdf

ISPs can leverage tracking done by non-ISPs for their own benefit

- When trackers are unencrypted, ISPs can leverage identifiers in tracking pixels/resource loads to disambiguate multiple users using the same IP address.
- Our 2014 study[1] concluded that network surveillance could analyze the relationships between first party web pages and the different third party trackers to infer a user’s browsing history when IP addresses aren’t enough.

[1] Englehardt, et al., Cookies That Give You Away: The Surveillance Implications of Web Tracking (2015) http://www.www2015.it/documents/proceedings/proceedings/p289.pdf

Users have fewer options against ISP data collection than they do against web-based data collection

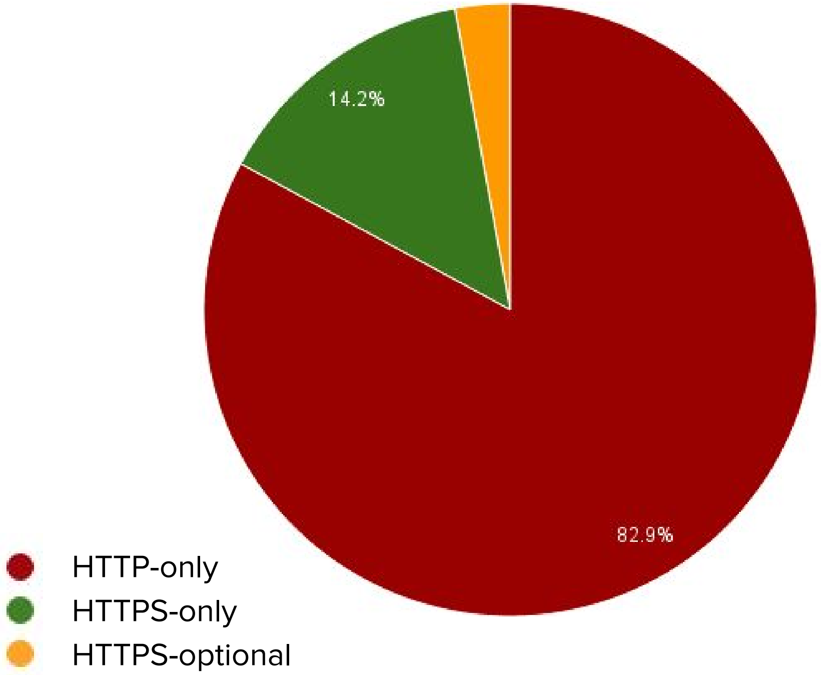
Tools to avoid non-ISP tracking and data collection work, but are not effective against ISPs

- We found web privacy tools (browser extensions like Ghostery, Adblock Plus, etc.) to be largely effective at blocking prominent third parties.
- But these tools do nothing to stop data collection on the wire.[1]

[1] Englehardt, S. and Narayanan, N., Online tracking: A 1-million-site measurement and analysis (2016), http://randomwalker.info/publications/OpenWPM_1_million_site_tracking_measurement.pdf

“To what extent does an end user have control over the use of encryption?”

Browser plugins that enable HTTPS when available will not help on many websites



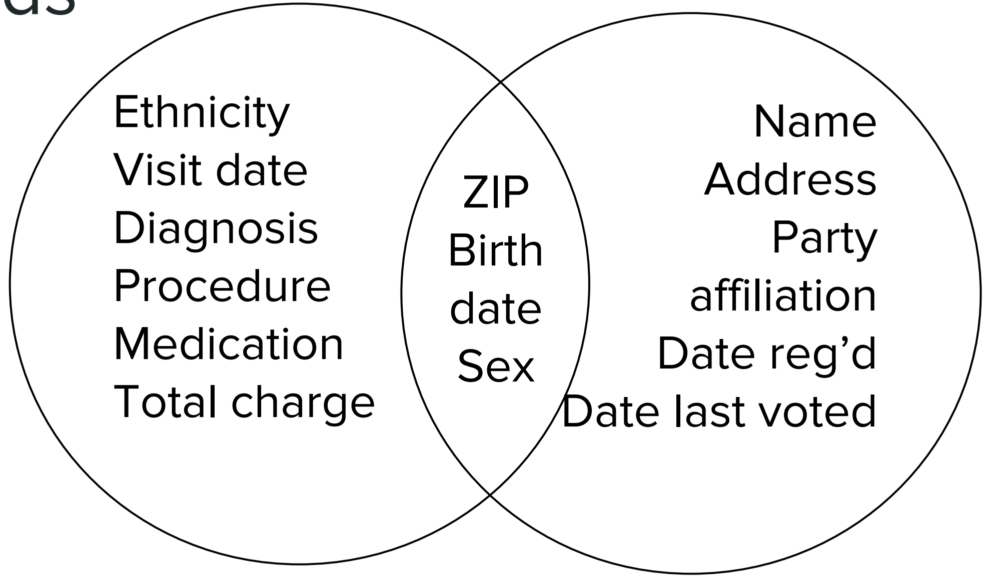
- Popular browser extension “HTTPS Everywhere” (>1 million users) forces browser to use HTTPS wherever possible.
- From our 2016 study[1], of the top 55,000 websites, **only 2.9% of websites that default to HTTP are also capable of HTTPS connections**.

[1] Englehardt, S. and Narayanan, N., Online tracking: A 1-million-site measurement and analysis (2016), http://randomwalker.info/publications/OpenWPM_1_million_site_tracking_measurement.pdf

Some thoughts on de-identification

Re-identification has a habit of surprising us

Latanya Sweeney’s re-identification of Mass. hospital records



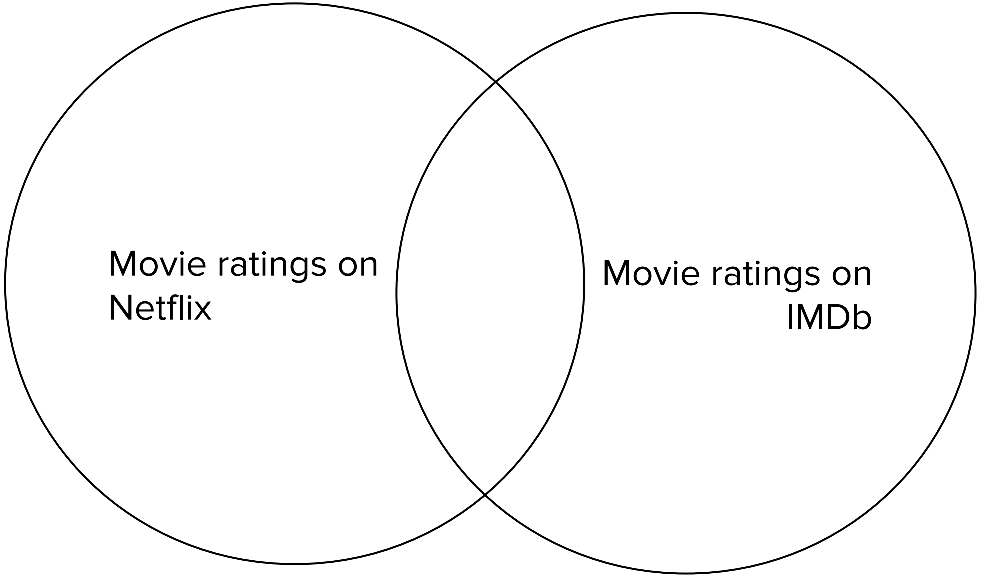
Ethnicity
Visit date
Diagnosis
Procedure
Medication
Total charge

ZIP
Birth date
Sex

Name
Address
Party affiliation
Date reg’d
Date last voted

Sweeney, L. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems (2002). https://epic.org/privacy/reidentification/Sweeney_Article.pdf

Re-identification of the Netflix Prize dataset



Narayanan, A., & Shmatikov, V. Robust de-anonymization of large sparse datasets. (2008)
https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf

Lesson from CS research:

Distinction b/w PII & non-PII is not useful as criterion for what is re-identifiable

Re-identifiability has been demonstrated repeatedly

● Location data	de Montjoye, Y., et al. "Unique in the crowd: The privacy bounds of human mobility." (2013).
● Credit card data	de Montjoye, Y., et al. "Unique in the shopping mall: On the reidentifiability of credit card metadata." (2015).
● Social network structure	Narayanan, A, and Shmatikov, V. "De-anonymizing social networks." (2009).
● Writing style	Narayanan, A., et al. "On the feasibility of internet-scale author identification." (2012).
● Programmers’ coding style	Caliskan-Islam, A., et al. "De-anonymizing programmers via code stylometry." (2015).
● Typing cadence	Monrose, F., & Rubin, A. D.. Keystroke dynamics as a biometric for authentication. (2000)
● Genetic data	Gymrek, M., et al. Identifying personal genomes by surname inference. (2013).

Lesson from CS research:

In the vast majority of cases, longitudinally linked data cannot be effectively anonymized

Example: source → destination IP logs

```
 13. 22.199. 62  →   248.171.115.104
 87.117.151.199  →   124. 64.221.231
180.240.243.169  →   181.177.121.204
249. 95. 74.142  →    34. 39.227. 82
173.103.202.180  →   248.171.115.104
180.240.243.169  →   107. 44. 58.251
180.240.243.169  →    44.154.213.249
211.158. 32.127  →    86. 14.198.117
```

Example: source → destination IP logs

```
 13. 22.199. 62  →   248.171.115.104
 87.117.151.199  →   124. 64.221.231
180.240.243.169  →   181.177.121.204
249. 95. 74.142  →    34. 39.227. 82
173.103.202.180  →   248.171.115.104
180.240.243.169  →   107. 44. 58.251
180.240.243.169  →    44.154.213.249
211.158. 32.127  →    86. 14.198.117
```

May reveal profile of websites visited by an individual.

Research suggests that such a profile is unique to the individual.

Cross-link with Twitter, IMDb, etc.

A precautionary approach

- Burden of proof should rest with the company/provider
    - -Provable privacy: encryption, differential privacy, ...
- Governments have many levers to incentivize
- Risk analysis should be qualitative, not quantitative

Narayanan, A., Huey, J., & Felten, E. W., A Precautionary Approach to Big Data Privacy (2016)
http://randomwalker.info/publications/precautionary.pdf

Recommendation: exceptions for longitudinally unlinked data and summary data

Key takeaways

1. Encryption is not yet pervasive enough to mitigate privacy concerns.
2. Unencrypted web traffic regularly contains sensitive and valuable customer data.
3. ISPs have unique and comprehensive access to users’ activities on the web.
4. De-identification has serious limitations.

Thank you!